UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Collection Dp

Rapports de Recherche

N° 1494

Programme 5
Traitement du Signal,
Automatique et Productique

TOWARDS A STATISTICAL THEORY OF INTENTIONS FOR KNOWLEDGE ANALYSIS

Edwin DIDAY

Juillet 1991

TOWARDS A STATISTICAL THEORY OF INTENTIONS FOR KNOWLEDGE ANALYSIS

VERS UNE THEORIE STATISTIQUE DES INTENTIONS POUR L'ANALYSE DES CONNAISSANCES

Edwin DIDAY
University Paris Dauphine /INRIA

Résumé

Le but principal de l'approche symbolique en statistique est d'étendre la problématique, les méthodes et les algorithmes utilisés sur des données classiques (points de $\mathbb{R}^n$, par exemple) à des données plus complexes caractérisées par des objets dits symboliques. Ces objets sont mieux adaptés à la représentation de connaissances car ils sont définis en intention en "unifiant" contrairement aux observations généralement traitées par la statistique qui caractérisent des objets "individuels". On introduit différentes sortes d'objets symboliques, booléens, possibilistes, probabilistes et reliés à la théorie de l'évidence. On résume ensuite quelques-unes de leurs qualités et propriétés. On donne quelques idées sur la façon dont la théorie des probabilités, la statistique et l'analyse des données peuvent être étendues à ces objets. Enfin quatre type de problèmes d'analyse des données incluant l'extension symbolique sont présentés.

Abstract

The main aim of the symbolic approach in statistics is to extend problems, methods and algorithms used on classical data to more complex data called "symbolic objects"which are well adapted to representing knowledge and which "unify" unlike usual observations which characterize "individual things". We introduce several kinds of symbolic objects : boolean, possibilist and probabilist, related to belief theory. We briefly present some of their qualities and properties. We give some ideas on how probability theory, statistics and data analysis may be extended on these objects. Finally four kinds of data analysis problems are presented.

Introduction

In probability theory , very little is said about events which are generally identified to parts of the set of samples Ω. In computer science, object oriented langages consider more general events called objects or "frames" defined by intention. In data analysis (multidimensional scaling, clustering, exploratory data analysis etc.) more importance is given to the elementary objects which belong to the sample Ω then in classical statistics where attention is focused on the probability laws of Ω ; however, objects of data analysis are generally identified to point of $\mathbb{R}^p$ and hence are unable to treat complex objects coming for instance from large data bases, and knowledge bases. Our aim is to define complex objects called "symbolic objects" inspired by those of oriented object langages in such a way that data analysis become generalized in knowledge analysis. Objects will be defined by intention by the properties of their extension. More precisely, we distinguish objects which "unify" rather than elementary observed objects which characterize "individual things" (their extension) : for instance "the customers of my shop" instead of "a customer of my shop", "a specie of mushroom" instead "the mushroom that I have in my hand".

The aim of this paper is to reduce the gap between statistics or data analysis (where people are not yet very interested in treating this kind of objects) and artificial intelligence (where people are more interested in knowledge representation, reasoning and learning then in knowledge analysis).

1. Boolean symbolic objects

We consider Ω a set of individual things called "elementary objects" and a set of descriptor functions $y_i : \Omega \rightarrow O_i$.

A basic kind of symbolic objects are "*events* ". An event denoted $e_i = [y_i=V_i]$ where $V_i \subseteq O_i$ is a function $\Omega \rightarrow$ {true, false} such that $e_i(w)$ = true iff $y_i(w) \in V_i$. When $y_i(w)$ has no sense (the kind of computer used by a company without computer) $V_i = \phi$ and when it has a meaning but it is not known $V_i = O_i$. The extension of e_i in Ω denoted by ext (e_i/Ω) is the set of elements $w \in \Omega$ such that $e_i(w)$ = true.

An assertion is a conjunction of events $a = \wedge_i [y_i=V_i]$; the extension of a denoted ext(a/Ω) is the set of elements of Ω such that $\forall i \; y_i(w) \in V_i$.

A "*horde* " is a symbolic object which appears for instance, when we need to express relations between parts of a picture that we wish to describe. More generally a horde is a function h from Ω^p in {true, false} such that $h(u) = \wedge_i [y_i(u_i) = V_i]$ if $u = (u_1,....,u_p)$. For example :

$h = [y_1(u_1) = 1] \wedge [y_2(u_2) = \{3,5\}] \wedge [y_3(u_1) = [30,35]] \wedge$ [neighbour (u_1,u_2) = yes].

A synthesis object is a conjunction or a semantic link between hordes denoted in case of conjunction by $s = \wedge_i h_i$ where each horde may be defined on a different set Ω_i by different descriptors. For instance Ω_1 may be individuals, Ω_2 location, Ω_3 kind of job etc. All these objects are detailed in Diday (1991).

2. Modal objects

Suppose that we wish to use a symbolic object to represent individuals satisfying the following sentence : "It is possible that their weight be between 300 and 500 grammes and their color is often red or seldom white" ; this sentence contains two events e_1 = [color = {red, white}] which lack the modes *possible* , *often* and *seldom* , a new kind of event, denoted f_1 and f_2, is needed if we wish to introduce them f_1 = *possible* [height = 300,500]] and f_2 = [color = {*often* red , *seldom* white}] ; we can see that f_1 contains an *external* mode *possible* affecting e_1

whereas f_2 contains *internal* modes affecting the values contained in e_2. Hence, it is possible to describe informally the sentence by a modal assertion object denoted $a = f_1 \wedge_x f_2$ where $\wedge_x$ represents a kind of conjunction related to the background knowledge of the domain. The case of modal assertions of the kind $a = \wedge_i f_i$ where all the f_i are events with external modes has been studied for instance in Diday (1990). This paper is devoted to the case where all the f_i contain only internal modes.

3. Internal modal objects

3.1. A formal definition of internal modal objects

Let x be the background knowledge and

. M^x a set of modes, for instance M^x = {often, sometimes, seldom, never} or $M^x = [0,1]$.

. $Q_i = \{ q_i^j \}_j$ a set of mappings q_i^j from O_i in M^x, for instance O_i = {red, yellow, green},

$M^x = [0,1]$ and q_i^j (red) = 0.1 ; q_i^j (yellow) = 0.3 ; q_i^j (green) = 1, where the meaning of the values 0.1, 0.3, 1 depends on the background knowledge (for instance q_i^j may express a possibility, see §4.1)

. y_i is a descriptor (the *color* for instance) ; it is a mapping from Ω in Q_i. Notice that in the case of boolean objects y_i was a mapping from Ω in O_i, and not Q_i.

Example : if O_i and M^x are chosen as in the previous example and the color of w is red then $y_i(\omega) = r$ means that $r \in Q_i$ be defined by r (red) = 1, r(yellow) = 0, r (green) = 0.

. $OP_x = \{ \cup_x, \cap_x, c_x \}$ where $\cup_x, \cap_x$ expresses a kind of union and intersection between subsets of Q_i and c_x (q_i) (sometimes denoted $\bar{q}_i$, the complementary of $q_i \in Q_i$).

Example : if $q_i^j \in Q_i$ and $Q_i^j \subseteq Q_i$

$$q_i^1 \cup_x q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2$$

$$q_i^1 \cap_x q_i^2 = q_i^1 \; q_i^2 \text{ where } q_i^1 q_i^2 (v) = q_i^1 (v) \; q_i^2 (v) \; ; \; c_x (q_i) = 1 - q_i$$

$$Q_i^1 *_x Q_i^2 = b(Q_i^1) *_x b(Q_i^2) \text{ where } *_x \in \{ \cup_x, \cap_x \} \text{ and}$$

$$b(Q_i^j) = \{ \cup_x q_i / q_i \in Q_i^j \} \text{ and } c_x (Q_i^j) = 1 - c_x (b(Q_i^j)).$$

This choice of OP_x is "archimedian" because it satisfies a family of properties studied by Shweizer and Sklar (1960) and recalled by Dubois et Prade (1988).

. g_x is a "comparison" mapping from $Q_i \times Q_i$ in an ordered space L^x.

Example : $L^x = M^x = [0,1]$ and $g_x (q_i^1 , q_i^2) = \langle q_i^1 , q_i^2 \rangle$ the scalar product

. f_x is an "aggregation" mapping from $P(L^x)$ the power set of L^x in L^x . For instance,

f_X ({L_1,...,L_n}) = Max L_i.

Let Y = {y_i} be a set of descriptors and V = {V_i} a set of subsets of Q_i such that $V_i = \{q_i^j\} \subseteq Q$. Now we are able now to give the formal definition of an internal objet (called "im" object).

Definition of an im assertion

Given OP_X, g_X and f_X, an im object is a mapping a_{YV} from Ω in an ordered space L^X denoted a $= \underset{i}{\wedge} [y_i = \{q_i^j\}_j]$ *such that if $\omega \in \Omega$ is described for any i by $y_i(\omega) = \{r_i^j\}$ then*

$$a_{yv}(\omega) = f_X(\{g_X(\underset{j}{\vee_X} q_i^j, \underset{j}{\vee_X} r_i^j)\}_i).$$

We denote by $\mathcal{A}_X$ the set of im objects associated to background knowledge x and φ the mapping from Ω in $\mathcal{A}_X$ such that $\varphi(\omega) = \omega^S = \underset{i}{\wedge_X} [y_i = y_i(\omega)]$.

3.2. Extension of im objects

There are at least two ways to define the extension of an im object a. The first consists in considering that each element $\omega \in \Omega$ is more or less in the extension of a according to its weight given by a (ω) ; in that case the extension of a denoted Ext (a/Ω) will be the set of couples
{(ω, $a(\omega)$) / $\omega \in \Omega$}. The second requires a given threshold α and then the extension of a will be Ext (a/Ω, α) = {(ω, $a(\omega)$) / $\omega \in \Omega$, $a(\omega) \geq \alpha$}.

3.3. Semantic of im objects

In addition to the modes several other notions may be expressed by an im object a :

a) Certitude : $a(\omega)$ is not true or false as for boolean objects but it expresses a degree of certitude.

b) Variation : it appears at two levels, in an im object denoted a = $\underset{i}{\wedge_X} [y_i = \{q_i^j\}_i]$; first in each q_i^j, for instance if y_i is the color and q_i^1 (red) = 0.5, q_i^1 (green) = 0.3 it means that there exists a variation between the individual objects which belong into the extension of a (for instance a specie of mushrooms) where some are red and others are green ; second, for given description y_i between the q_i^j (each q_i^j expresses for instance the variation in a different kind of specie).

c) Doubt : if we say that the color of mushrooms of a specie is red "or" green, it is an "or" of variation, but if we say that the color of the mushroom which is in my hand is red "or" green, it is an "or" of doubt.
Hence, if we describe $\omega \in \Omega$ by $\varphi(\omega) = \omega^S = \underset{i}{\wedge}$ [$y_i = y_i(\omega)$] where $y_i(\omega) = \{r_i^j\}_j$ we express a doubt in each r_i^j and among the r_i^j provided for instance by several experts.

3.4. An example of background knowledge expressing "intensity".

Here the background knowledge x is denoted i for intensity. Each individual object $\omega \in \Omega$ is a manufactered object described by two features y_1 which express the degree of "roundness" and y_2 the "heaviness" : O_1 = {flat, round}, O_2 = {heavy} ; M^i = {*very, enough, a little, very little, nil*}

Let a and ω^S be defined by :

$a = [y_1 =$ *a little* flat, *enough* round$] \wedge_i [y_2 =$ *a little* heavy$]$

$\omega^S = [y_1 =$ *enough* round$] \wedge_i [y_2 =$ *a little* heavy, *enough* heavy$]$.

(The user has a doubt for ω between *a little* and *enough* heavy).

Hence q_1^1 (flat) = *a little* ; q_1^1 (round) = *enough* ; q_2^1 (heavy) = *a little* , r_1^1 (flat) = *nil* ; r_1^1 (round) = *enough* ; r_2^1 (heavy) = *a little* , r_2^2 (heavy) = *enough* .

A given taxonomy T which expresses the background knowledge on the values of M^i allows to say that Tax (*a little, very little*) = *rather* ; hence if we settle that

$r_2^1 \cup_i r_2^2 (v) =$ Tax $(r_2^1 (v), r_2^2 (v))$ we have $r_2^1 \cup_i r_2^2$ (heavy) = Tax (*a little, enough*) = *almost.*

We define L^i by L_1 = not acceptable, L_2 = acceptable, L_3 = completely acceptable and, we suppose that the comparison mapping g_i is given by a table T_g such that $g_i (q_1^1, r_1^1) = T_{g_i}$ ((*a little* flat, *enough* round), (*nil* flat, *enough* round)) = acceptable and
$g_i (q_2^1, r_2^1 \cup_X r_2^2) = T_{g_i}$ (*a little* heavy, *almost* heavy) = not acceptable.

Finally if we settle $f(\{L_i\}) = \text{Min } L_i$ and $L_1 < L_2 < L_3$ we obtain $a(\omega) = f_i$ (not acceptable, acceptable) = not acceptable.

4. Possibilist objects

4.1. The possibilist approach

Here we follow Dubois and Prade (1984) in giving the main idea of this approach.

Definition of a measure of possibility and of necessity

This is a mapping Π from $P(\Omega)$ the power set of Ω in [0, 1] such that

(1) $\Pi(\Omega) = 1 \qquad \Pi(\phi) = 0$

(2) $\forall A, B \subseteq \Omega \qquad \Pi(A \cup B) = Max\,(\Pi(A), \Pi(B))$

A measure of necessity is a mapping from $P(\Omega)$ in [0, 1] such that :

(3) $\forall A \subseteq \Omega \qquad N(A) = 1 - \Pi(\bar{A})$.

The following properties may then be shown :

$N(\phi) = 0$; $N(A \cap B) = \text{Min}\ (N(A), N(B))$; $\Pi\ (\cup_i A_i) = \underset{i}{\text{Max}}(\Pi(A_i))$; $N(\underset{i}{\cap} A_i) = \underset{i}{\text{Min}}$ $(N(A_i))$; $\Pi(A) \leq \Pi(B)$ if $A \subseteq B$; $\text{Max}\ (\Pi(A), \Pi(\bar{A})) = 1$; $\text{Min}\ (N(A), N(\bar{A})) = 0$; $\Pi(A) \geq N(A)$; $N(A) > 0$ implies $\Pi(A) = 1$; $\Pi(A) < 1$ implies $N(A) = 0$; $\Pi(A) + \Pi(\bar{A}) \geq 1$ and $N(A) + N(\bar{A}) \leq 1$.

Example

We define $\Pi_E(A)$ (resp. $N_E(A)$) as the possibility (resp. the necessity) to get $\omega \in A$ when $\omega \in E$. We say that $\Pi_E(A) = 1$ if this possibility is true and $\Pi_E(A) = 0$ if no. Hence Π_E and N_E are mapping from $P(\Omega)$ in $\{0, 1\}$. It is easy to show that Π_E and N_E satisfy the three axioms.

The theory of possibility models several kind of semantics, for instance :

i) The physical possibility : this expresses the material difficulty for an action to occur :"I have the possibility of carrying 20kg".

ii) The possibility as a concordance with an actual knowledge "it is possible that it will rain today".

iii) The non-astonishement : for instance, "the typicality" for the color of a flower to be yellow".

4.2. A formal definition of possibilist objects

Here the background knowledge x is denoted p as possibility.

Definition

A possibilist assertion denoted $a_p = \underset{i}{\wedge} [y_i = \{q_i^j\}_j]$ *is an im assertion which takes its values in* $L^P = [0, 1]$ *such that*

. $\forall i\ Q_i$ *is a set of measures of possibility.*

. OP_p : $\forall i, q_i^1, q_i^2 \in Q_i\ q_i^1 \cup_p q_i^2 = Max(q_i^1, q_i^2)$; $q_i^1 \cap_p q_i^2 = Min(q_i^1, q_i^2)$;

$c_p(q) = 1 - q$.

. g_p : $g_p(q_i^1, q_i^2) = sup\{min(q_i^1(v), q_i^2(v)) / v \in O_i\}$

. f_p : $\forall L \subseteq [0, 1]\ f_p(L) = Min(\ell / \ell \in L)$

Notice that OP_p is defined as in fuzzy sets and g_p has also been proposed by Zadeh (1971).

It is also possible to define a "necessitist" assertion a_n by setting $\forall \omega \in \Omega\ a_n(\omega) = 1 - a_p$ $(c(\omega))$ where $c(\omega) = \varphi^{-1}(c(\omega^S))$ and $c(\omega^S) = \underset{ip}{\wedge} [y_i = c(r_i)]$ if $\omega^S = \underset{ip}{\wedge} [y_i = r_i]$.

This results in $a_n(\omega) = 1 - f(\{g_p(q_i, \bar{r}_i)\}_i)$ and then $a_n(\omega) = 1 - \underset{i}{\text{Max}}\ g_p(q_i, \bar{r}_i)$

$= 1 - \max \{\sup \{\min (q_i(v), \bar{r}_i(v) / v \in O_i\}\}_i$

$= \min \{1 - \{\sup \min (q_i(v), \bar{r}_i(v)) / v \in O_i\}\}_i$

$= \min \{\inf \{1 - \min (q_i(v), \bar{r}_i(v)) / v \in O_i\}\}_i$

$= \min \inf \{\max (1 - q_i(v), r_i(v)) / v \in O_i\}$

and then finally $a_n(\omega) = \min g_n(q_i, r_i)$.

It results that a necessitist object is defined by $OP_n = \{\cup_n, \cap_n, c_n\}$ where $\cup_n$ is $\cap_p$, $\cap_n$ is $\cup_p$ and c_n is c_p, $g_n(q_i, r_i) = \inf\{\max(\bar{q}_i(v), r_i(v)) / v \in O_i\}$ and $f_n = \min$.

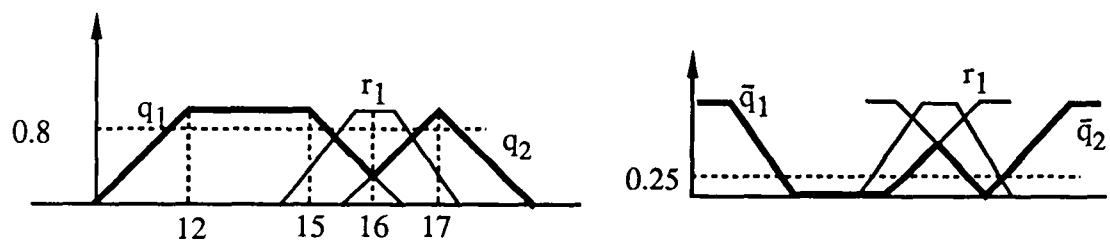
Example

An expert describes a class of objects by the following possibilist assertion (restricted, to simplify, to a single event) :

e_p = [height = [around [12, 15], about {17}]]. An elementary object ω is defined by ω^S = [height = close from 16].

The question is to find the possibility and necessity of ω knowing e_p, in the case where e_p and ω^S may be written :

e_p = [height = q_1, q_2] and ω^S = [height = r_1] where q_1, q_2, r_1 are possibilist mappings from O = [0, 20] in [0, 1] defined by the background knowledge on figure 1. This means that an object of height 14 (resp. 6) has a possibility 1 (resp. $\frac{1}{2}$). It is then possible to compute the possibility of ω by :



(a) (b)

Figure 1
(a) $q_1 \cup q_2 = \text{Max}(q_1, q_2)$
(b) $\bar{q}_1 \cap \bar{q}_2 = \text{Min}(\bar{q}_1 \cap \bar{q}_2)$

$e_p(\omega) = g_p(q_1 \cup_p q_2, r_1) = \sup\{\min(q_1 \cup_p q_2(v), r_1(v)) / v \in \Omega\} = 0.8.$

The necessity of ω is given by :

$$e_n(\omega) = g_n(q_1 \cup_p q_2, r_1) = \inf\{\max(\overline{q_1 \cup_p q_2}(v), r_1(v)) / v \in 0\} = 0.25.$$

This example shows that possibilist objects are able to represent not only certitude, variation and doubt but also vagueness (around, about) and inacuracy (close from 16).

5. Probabilist objects

5.1. The probabilist approach

First we recall the well known axioms of Kolmogorov :

If $C(\Omega)$ is a σ-algebra on Ω (i.e. a set of subsets stable for numerable intersection or union and for complementary).We say that p is a measure of probability on $(\Omega, C(\Omega))$ if

i) $p(\Omega) = 1$

ii) $p(\cup_i A_i) = \sum p(A_i)$ if $A_i \in C(\Omega)$ and $A_i \cap A_j = \phi$.

There are several semantics which follow these axioms : for instance luck in games, frequencies, some kind of uncertainty. Let Q_i be a set of measure of probabilities defined on $(O_i, C(O_i))$.

Definition

A probabilist assertion is an im assertion which takes its values in $L^{pr} = [0,1]$

$OP_{pr} : \forall q_i^1, q_i^2 \in Q_i \quad q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2 ; q_i^1 \cap_{pr} q_i^2 = q_i^1 q_i^2$ *which is the mapping which associate at* $v \in O_i$, $q_i^1(v) q_i^2(v)$;

$g_{pr} : \forall q_i^1, q_i^2 \in O_i \quad g_{pr}(q_i^1, q_i^2) = < q_i^1, q_i^2 > = \sum \{q_i^1(v) q_i^2(v) / v \in \Omega\}$.

$f_{pr} : f_{pr}(\{L_i\})$ = *mean of the* L_i.

To give an intuitive idea of the notion of union of measure of probabilities it is easy to see that if q_i^1 and q_i^2 are the measure of probabilities associated to two dices, $q_i^1 \cup_{pr} q_i^2$ (V) is the probability that the event V occurs, when the two dices are trialed independantly, in one dice or (not exclusive) in the other. Notice that $q_i^1 \cup_{pr} q_i^2$ is not a measure of probability because even if $q_i^1 \cup_{pr} q_i^2$ (v) $\in$ [0,1] the sum of the $q_i^1 \cup_{pr} q_i^2$ (v) on O_i is larger than 1.

5.2. Example

Un object ω is described by its color $y_1(\omega)$ which may be red or blue and its roundness $y_2(\omega)$ which may be round or flat.

Let $a = [y_1 = q_1^1, q_1^2] \wedge_{pr} [y_2 = q_2]$ and $\omega^s = [y_1 = r_1] \wedge_{pr} [y_2 = r_2]$ where q_1^1 (red) = 0.9 ; q_1^1 (blue) = 0.1 ; q_1^2 (red) = 0.5 ; q_1^1 (blue) = 0.5 ; q2 (round) = 0.2 ; q2 (flat) = 0.8. It results that a is described by two kind of objects : either often red and rarely blue or red or blue with same probability.

By using $q_1^3 = q_1^1 \cup_{pr} q_1^2 = q_1^1 + q_1^2 - q_1^1 q_1^2$ we obtain

q_1^3 (red) = 0.9 + 0.5 - 0.9 x 0.5 = 0.95

q_1^3 (blue) = 0.1 +0.5 - 0.1 x 0.5 = 0.55

We have also :
r_1 (red) = 1, r_1 (blue) = 0; r_2 (round) = 1, r_2 (flat) = 0 and then it results that

$a(\omega) = g_{pr}(q_1^3, r1) \wedge_{pr} g_{pr}(q_2, r_2)$

$= (0.95 \times 1 + 0.55 \times 0) \wedge_{pr} (0.2 \times 1 + 0.8 \times 0)$

$= 0.95 \wedge_{pr} 0.20 = (0.95 + 0.20) = 0.57$, which represents a membership degree of ω to the im object defined by a.

5.3. Kernel objects, credibility and plausibility

Given q_i a measure of probability on $(O_i, C(O_i))$ and we denote $q_i^j(V) = q_i(V_j \cap V)$ and we suppose that the V_j are chosen such that $\sum_j q_i^j(V_j) = 1$. The q_i^j for a given set of $V_j \subseteq O_i$ are called kernels and we denote $Q_i = \{q_i^j\}_j$

Definition

A kernel object denoted $a_k = \hat{\wedge}_{ik} [y_i = \{q_i^j\}_j]$ *is an im assertion which takes its values i* $L^k = [0,1]$ *such that :*

. $OP_k : \forall q_i^1, q_i^2 \in Q_i \qquad q_i^1 \cup_k q_i^2 (V) = q_i((V_1 \cup V_2) \cap V)$

$q_i^1 \cap_k q_i^2 (V) = q_i((V_1 \cup V_2) \cap V))$ *and* $c_k(q_i^j)(V) = q_i(c(V_j) \cap V)$.

. $g_k : g_k(q_i^1, q_i^2) = q_i(V_1 \cap V_2)$.

. f_k *is the mean.*

It results from this definition that if $a_k = \hat{\wedge}_{ik} [y_i = \{q_i^j\}_j]$ given $\omega \in \Omega$ such that

$\omega^s = \hat{\wedge}_{ik} [y_i = r_i^\ell]$ we have $a_k(\omega) = g_k(\{q_i^j\}_j, r_i^\ell) = q_i(\bigcup_j V_j \cap U_\ell)$ if $r_i^\ell(V) = q_i(U_\ell \cap V)$.

In practice it may happen that we are able to know only the q_i^j and not q_i ; in this case we may compute $a_k(\omega)$ by approximating $q_i((\bigcup_j V_j) \cap U_\ell)$ with a measure denoted E_α such that

$E_i^{\alpha,\beta}(V) = \sum_j \{q_i^j(V) / d^\circ(V_j \cap V) \in I_{\alpha,\beta}\}$

where $d^\circ(V_j \cap V)$ expresses a degree of intersection between V_j and V and
$I_{\alpha,\beta} = [\alpha, \beta] \subseteq [0,1]$ expresses for instance that the intersection of V_j with V is between 100α % and 100β %. If $\alpha=\beta=1$ V_j is included in V and $E_i^{1,1}(V) = Bel(V)$ if $\alpha=\beta=0$, $E_i^{0,0}$

(V) = Pl (V) and Bel and Pl satisfy the axioms of belief and plausibility measures (see for instance Schafer (1976)), Pearl (1990)), $\sum_j q_i^j (V_j) = 1$, $Bel_i (V) = \sum \{q_i^j (V_j) / V_j \subseteq V\}$, $Pl_i (V) = 1 - Bel (c (V))$. It is also easy to see that $Bel < q_i < Pl$.

6. The particular case of boolean objects

A boolean object $a = \hat{\wedge}_i [y_i = V_i]$ is an im object $a_b = \hat{\wedge}_i [y_i = q_i]$ where q_i is the characteristic mapping of V_i in O_i, $OP_b = \{\cup_b, \cap_b, c_b\}$ is such that $q_1 \cup_b q_2 = Max(q_1, q_2)$, $q_1 \cap_b q_2 = \min(q_1, q_2)$ et $c_b (q) = 1 - q$;if $w = \hat{\wedge}_i [y_i = r_i]$ where r_i is the characteristic mapping of $y_i(\omega)$ in O_i, $g_b(r_i, r_i) = <q_i, r_i>$ and $f_b = \min$; it results that if it exists only a single $v \in O_i$ such that $r_i (v) \neq 0$ then $a_b(\omega) = 1$ (thus $r_i \leq q_i$) $\Leftrightarrow a(\omega) =$ true and then $a_b(\omega) = 0 \Leftrightarrow a(\omega) =$ false.

7. Some qualities and properties of symbolic objects

7.1. Order, union and intersection between im objects

It is possible to define a partial preorder $\leq_\alpha$ on the mi objects by setting that : $a_1 \leq_\alpha a_2$ iff $\forall \omega \in \Omega \ \alpha \leq a_1 (\omega) \leq a_2 (\omega)$.

We deduce from this preorder an equivalence relation R by $a_1 R a_2$ iff $Ext (a_1 / \Omega, \alpha) = Ext (a_2 / \Omega, \alpha)$ and a partial order denoted $\leq_\alpha$ and called "symbolic order" on the equivalence classes induced from R.
We say that a_1 inherits from a_2 or that a_2 is more general than a_1 , at the level α, iff
$a_1 \leq_\alpha a_2$ (which implies $Ext_\alpha (a_1 / \Omega, \alpha) \subseteq Ext_\alpha (a_2 / \Omega, \alpha)$).
The symbolic union $a_1 \cup_x a_2$ (resp. intersection $a_1 \cap_x a_2$) at the level α is the conjunction $\wedge_x$ of the im objects b such that $Ext (a_1 / \Omega, \alpha) \cup Ext (a_2 / \Omega, \alpha) \subseteq Ext (b / \Omega, \alpha)$ (resp. $Ext (a_1 / \Omega, \alpha) \cap Ext (a_2 / \Omega, \alpha) \subseteq Ext (b / \Omega, \alpha)$).

7.2.Some qualities of symbolic objects

As in the boolean case, see Brito, Diday (1989), it is possible to define different kinds of qualities of symbolic objects (refinement, simplicity, completeness ...).

For instance, we say that a symbolic object s is complete iff the properties which characterize its extension are exactly those whose conjunction defines the object. More intuitively if I see white dogs in my street and I state "I see dogs", my statement doesn't describe the dogs in a complete way, since I am not saying that they are white.

On the other hand,the simplicity at level α of an im object is the smallest number of elementary events whose extension at level α coincides with the extension of s at the same level.

7.3.Some properties of im objects

It may be shown see Diday (1991) for instance, that given a level α the set of im objects is a lattice for the symbolic order and that the symbolic union and intersection define the supremum and infimum of any couple. It may also be shown that the symbolic union and intersection of complete im objects are complete im objects and hence that the set of complete im objects is also a lattice.

8. An extension of probability theory

The notions of σ–algebra, measure, random variable and the Kolmogorov axioms may be extended (see Diday 91) in at least two ways

i) the set of samples Ω will be the set of im objects a_x

ii) the usual operators OP = {∪, ∩, c} on classical set theory will be replaced by $OP_x = \{\cup_x, \cap_x, c_x\}$ adapted to the background knowledge.

More precisely we get the following extensions associated with a background knowledge (b.k.) x :

Definition of a σ–algebra associated with b.k. x

A σ–algebra $C_x(E)$ on a set E provided with $OP_x = \{\cup_x, \cap_x, c_x\}$ is a set of subsets of E such that :

i) $E \in C_x(E)$

ii) $\forall E_i \in C_x(E), c_x(E_i) \in C_x(E)$

iii) any enumerable sequence $\{E_i\}$ *of subsets of E is such that* $\cap_{i x} E_i \in C_x(E)$ *an* $\cap_{i x} E_i \in C_x(E)$.

Definition of a probability measure associated with a b.k. x :

A probability measure q_x *on* $(E, C_x(E))$ *is a mapping* q_x *from* $C_x(E)$ *in [0,1] such that :*

i) $q_x(E) = 1$

ii) $\forall E_1, E_2, \in C_x(E), q_x(E_1 \cup_x E_2) = q_x(E_1) + q_x(E_2) - q_x(E_1 \cap_x E_2)$

Definition of a random variable X associated with a b.k. x:

A random variable X from $(E, C_x(E), q_x)$ *in* $(F, C_x(F))$ *is a mapping from* $(E, C_x(E))$ *in* $(F, C_x(F))$ *such that* $\forall F_i \in C_x(F), X^{-1}(F_i) \in C_x(E)$.

Such a random variable induces a mapping called the "law of X" denoted q_x^X, which is a mapping from $C_x(F)$ in [0,1] such that $q_x^X(F_i) = q_x(X^{-1}(F_i))$.

It may then be shown that if $X^{-1}(F_i *_x F_j) = X^{-1}(F_i) *_x X^{-1}(F_j)$ for $*_x \in \{\cup_x, \cap_x\}$ then the law of X is a probability measure on $(E, C_x(E))$.

We use those definitions to extend an im assertion $a = \hat{\wedge}_i [y_i = q_i]$ to a dual im assertion denoted a* defined on subsets of a_x the set of im assertions. More precisely :

Notice that it may be shown in the case of kernel objects that OP_x is an idempotent algebra whereas in case of probabilist objects, OP_x is archimedian.

Given $A \subseteq a_x$ we denote a^*_ℓ a "dual" measure of $a_\ell = \hat{\wedge}_{ix} [y_i = q_i^\ell]$ and Q_i^A the set of q_i^j such that $a = \hat{\wedge}_{ix} [y_i = q_i^j] \in A$, and we settle

$$a^*_\ell(A)=f_x(\{g_x(q_i^\ell, \{\cup_x^j q_i^j / q_i^j \in Q_i^A\}_i))$$

Then, it may be shown that a* is the law of a random variable X from $(Q_x, C_x(Q_x), q_x^*)$ in $(a_x, C_x, a_x))$ where q_x^* is a probability measure defined on Q_x, the set of $q^\ell = \{q_i^\ell\}_i$ mappings. Hence, in the case of probabilist or kernel objects we get :

i) $a^*(a_x) = 1$ and ii) $\forall A_1, A_2 \subseteq a_x \quad a^*(A_1 \cup_x A_2) = a^*(A_1) + a^*(A_2) - a^*(A_1 \cap_x A_2)$.

Finally it appears that a* is a kind of probability on probabilist objects in case of probabilist objects.

In the case of possibilist objects it may be shown that a* is kind of possibility on possibilist objects.

i) $a^*(a_p) = 1$ and

ii) $\forall A_1, A_2 \subseteq a_p \qquad a^*(A_1 \cup_p A_2) = \text{Max}\ a^*(A_1),\ a^*(A_2))$.

9. Statistics and data analysis of symbolic objects

Several works have been recently carried out in this field : for histograms of symbolic objects, see De Carvalho & al (1990) and (1991) ; for generating rules by decision graph on im objects in the case of possibilist objects with typicalities as modes see Lebbe and Vignes (1991) ; for generating overlapping clusters by pyramids on symbolic objects see Brito, Diday (1990).

More generally, four kinds of data analysis may roughly be defined depending on the input and output : a) numerical analysis of classical data tables b) numerical analysis of symbolic objects (for instance by defining distances between objects) c) symbolic analysis of classical data tables, for instance obtaining a factor analysis or a clustering automatically interpreted by symbolic objects d) symbolic analysis of symbolic objects where the input and output of the methods are symbolic objects.

Conclusion

Unlike most of work carried out in expert systems, symbolic data analysis constitutes a "critique of pure reasoning" by giving less importance to the inference engine and more importance to the study of the knowledge base, considered as a set of "symbolic objects".
A wide field of research is opened by extending classical statistics to statistics of intensions and more specialy by extending problems, methods and algorithms of data analysis to symbolic objects.

References

. Brito P., Diday E., (1990), "*Pyramidal representation of symbolic objects*", in NATO ASI Series, Vol. F 61 Knowledge Data and computer-assisted Decisions edited by Schader and W. Gaul. Springer Verlag.

. De Carvalho F.A.T. (1991), "Histogramme en Analyse des Données Symboliques", Rapport de Recherche INRIA (à paraître).

. Diday E., (1990), "*Knowledge representation and symbolic data analysis*", in NATO ASI Series, Vol. F 61 Knowledge Data and computer-assisted Decisions edited by Schader and W. Gaul. Springer Verlag.

. Diday E., (1991), "*Objets modaux pour l'analyse des connaissances*", Rapport INRIA, Rocquencourt, 78150, France.

. Dubois D., Prade H., (1988), "*Possibility theory*", Plenum New York.

. Lebbe J., Vignes R., Darmoni S., (1990), "*Symbolic numeric approach for biological knowledge representation : a medical example with creation of identification graphs*", in : Proc. of Conf. on Data Analysis, Learning Symbolic and Numerical Knowledge, Antibes ed. E. Diday, Nova Science Publishers, Inc., New York.

. Schafer G., (1976), "A Mathematical Theory of evidence" Princeton University Press.

. Zadeh L.A.(1971), "*Quantitative fuzzy semantics*", Information Sciences, 159-176.

Imprimé en France
par
. l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249 - 6399